Complex Systems **3** (1989) 9–28

A Travelling Salesman Approach to Protein Conformation

Henrik Bohr
Søren Brunak
Department of Structural Properties of Matter,
The Technical University of Denmark, DK-2800 Lyngby, Denmark

Abstract. A simple method for finding conformational substates of proteins is presented and realized through computer simulations. It is based on a procedure in which the amino acids in a protein take the places of the cities in the three-dimensional travelling salesman problem. Optimization by simulated annealing was employed in the computer simulations to obtain conformational substates originating from a given three-dimensional structure of the protein backbone. Two polypeptides, Avian Pancreatic Polypeptide and Leucin-Enkephalin, were modelled and compared with available x-ray diffraction data. The method gives an interesting spinoff: the possibility of assigning a measure of complexity to real protein structures, due to the fact that a metric on the set of interactions employed by the protein can be defined naturally.

1. Introduction

The determination of the tertiary structure of proteins from their sequence of amino acids, and the functional role of conformational substates of a given tertiary state are important unsolved problems in the science of molecular biology. The dynamical behavior of complex systems like proteins is in general extremely hard to simulate on a computer. It has been suggested that the reason might be that many complex systems actually function as computers, performing computations which cannot be completed in fewer logical steps than the systems are using themselves. In other words, the computations are irreducible and the computational results cannot be obtained by the use of short cuts, because they do not exist [1].

In the case of the dynamics of proteins, where parts of the proteins move relatively to each other, the question is: what is the protein computing while it moves, and what is the final result of the computation?

The result of the protein folding process is clearly a topology. The process looks for good neighbors for the various parts, and the system ends up in a conformational state with high stability in which parts with an affinity

for each other (or a dislike of the solvent) are brought together, and where mutually repulsive parts seek positions with the highest possible distance between them. Equilibrium fluctuations make it possible for a protein in a given conformational state to assume a large number of similar substates [2,3], and thus the computational result must be described in the form of a time-dependent topology, or simply as an ensemble of the substates assumed.

In this process of sampling the set of topologies, the ratio between attraction and repulsion is often rather large, as can be seen from the very compact objects the folding process is capable of producing.

The determination of the three-dimensional structure of these folded objects by conventional molecular dynamics [4–8] leads to representations of large sets of coupled differential equations describing the interaction and motion of each atom in the protein. However, even with the help of today's supercomputers, an extreme computational complexity is connected with this application of brute force. Therefore much research activity has recently focused on simple models based on analogies to systems in physics with certain generic properties overlapping with those of the proteins, but which are better understood [9,10].

Proteins can be viewed roughly as being folded strings to which a number of side-chains are attached. The multiple configurations of these side-chains and their stability is the subject we consider in this paper. Important characteristics of protein dynamics are the many local minima in the energy surfaces, giving rise to many different metastable substates separated by barriers. These characteristics have also been found in disordered and frustrated systems [11] such as spin glasses, neural networks and in complex problems in optimization, of which the travelling salesman problem is a prototypical example.

The hill-valley energy profiles for proteins and the transition scheme of slow sequential relaxation, suggest that proteins in many ways are similar to glasses [9–11]. Systems consisting of simple parts behave in general with a degree of complexity which is intimately related to the diversity of the interactions present in the systems [12]. Examples are ferromagnets and spin glasses, of which the former employs very few different interactions and has a relatively simple behavior compared with the latter, which employs a wealth of different interactions and has a rather complex behavior. In systems of intermediate size, as proteins typically are, the interaction diversity will usually be very high [12].

Conformational substates of a protein are partly determined by the positions of the side-chains, and in a protein which has N side-chains, there are $(N-1)!/2$ side-chain—side-chain interactions to account for. This is clearly computationally intractable for larger proteins, but since the side-chain—side-chain interaction is predominantly of short range, we shall consider a model where only neighbor interactions are included. We are then left with the problem of choosing a suitable neighbor topology for the three-dimensional configuration of side-chains, being the result of a global relaxation throughout the protein. A suitable selection of neighbors should com-

prise short range as well as interactions of longer range. We will propose to achieve this by expanding the model so that it incorporates a procedure for selecting a self-consistent sequence of nearest neighbors. The most reasonable number of nearest neighbors may well differ from two. But for reasons of simplicity, we assume that the coordination number is fixed and equal to two. This is the smallest number which assures that the dynamics of the side-chain configuration is collective. As described later, it also immediately leads us to the analogy with the problem of the travelling salesman. The side-chains or atoms take the places of the cities, and the Hamiltonian becomes the sum of the neighbor interactions.

The Travelling Salesman Problem (TSP) is a finite problem in which we also encounter an energy surface with many local minima [13,14]. The TSP, when applied to protein structure, has the advantage of giving a simple dynamical description of protein conformations in terms of the kinematics of the participating side-chains. At the same time, it is easy to develop an optimization procedure on a TSP that, when transferred to the protein structure, gives a protein substate or a near-optimal local minimum of the Hamiltonian. In other words, our goal is to describe the set of functional substates and the involved interactions, based on the participating atoms movements, without invoking the full set of coupled differential equations of the motion for all atoms in the protein.

The function of a protein can be described either by the ensemble of substates parameterised by conformational coordinates, or by the ensemble of interactions which are responsible for the set of substates. The internal structure of these sets, i.e. the interrelations between set-members, possibly that of a hierarchy, as it is suggested for the first ensemble in reference [9], can be investigated. If a hierarchical structure is found, it is possible, by means of the complexity measure introduced in reference [12], to assign measures of *functional complexity* to proteins, and thus it becomes possible to relate the sequence of amino acids to the function of the protein.

In the following sections we shall further elaborate on the analogy between protein dynamics and the TSP. In section 2, TSP is introduced, and in section 3 it is applied to proteins. Finally in section 4 we present data from computer simulations of the model and compare them with x-ray data of the native structure. An appendix contains an outline of a procedure which assigns a measure of functional complexity to real protein structures.

2. The travelling salesman problem and optimization by simulated annealing

The travelling Salesman Problem is a complex combinatorial optimization problem [15] that is hard to solve, but easy to formulate. It belongs to the class of "NP-complete" problems [16] for which known algorithms, obtaining the exact solution, require a number of steps that grows at least exponentially with the number of constituents in the problem. The TSP is simply stated as the following. Given N cities and their mutual distances, the objective is

to find a tour, or a permutation P, of the cities such that the total length L of the tour, visiting all cities once and returning to the first in the succession P, is minimized. If d_{ij} denotes the distance between city i and j, the total length to be minimized is simply:

$$L = \sum_{i=1}^{N} d_{P(i)P(i+1)} \tag{2.1}$$

with the return condition:

$$P(N+1) = P(1) \tag{2.2}$$

If the starting point and direction of the tour do not matter, the total number of distinct tours is $(N-1)!/2$. To find the exact solution to the TSP by testing all possibilities is therefore impossible, if N is large. Heuristic strategies, which search for near optimal tours and only explore the phase space locally, will in most cases fail to find the overall optimal tour. This is because it often pays to choose cities that are not nearest neighbors at a given step of city-search, in order to optimize the total length. This phenomenon, which makes the TSP optimization very intricate, is termed frustration [11], and is, as in the case of spin glasses, the global inability to satisfy simultaneously all local ordering requirements. It is the cause of the many local minima of the energy, which here is equivalent to the length of the total travelling distance.

Simulated annealing [17,18] is a Monte Carlo optimization scheme designed to cope with these problems. It is based on the Metropolis algorithm [19] which accepts not only down-hill changes of the objective function or Hamiltonian, but also steps that raise it. In the approach a random number generator is used to generate new configurations.

In the TSP, a new configuration is made by introducing a number of n-bond moves into the old one. For a 2-bond move, two links on a tour are discarded, say $d_{i,i+1}$ and $d_{j,j+1}$, and then replaced by $d_{i,j}$ and $d_{i+1,j+1}$, so that the new path is again a tour. There are $N(N-1)/2$ such moves arising from any tour.

This procedure is performed with the given Boltzmann acceptance probability a number of times, forming a Markov chain, and eventually the system reaches thermal equilibrium with the macroscopic parameters fluctuating about their mean values in a Boltzmann distribution appropriate to the temperature. The simulated annealing procedure starts with a sufficiently high temperature, at which a relatively large amount of tour proposals are accepted, and the system can be considered as being melted. After waiting for equilibrium to be well established at each temperature, the temperature is decreased according to a cooling schedule. Finally, at a temperature where the system has frozen into a certain configuration, one hopes to have reached a 2-optimal solution, which is a solution stable to all 2-bond moves, and hopefully close to the global minimum or ground state of the problem.

3. TSP applied to protein conformations

3.1 The relaxation

The goal of this paper is to obtain a relaxation scheme for realistic spatial configurations of the atoms in a given protein. The relaxation we have in mind is based on an underlying hierarchically constrained dynamics, as described in reference [20] for a glassy relaxation, and which we believe is also essential for proteins. The basic feature of a hierarchically constrained dynamics is that the relaxation follows a series of many correlated activation steps. Consider a complex system with a series of levels each containing a number of degrees of freedom, and where each level is strongly correlated to the preceding level. In a discrete series of levels, $l = 0, 1, 2, \ldots$, the entities in level $l + 1$ are only free to change their state if a condition on some entities in level l is satisfied, a situation bearing a resemblance to a traffic jam.

For a protein molecule in the tertiary structure we ask what global effects a change in the position of one side-chain might have on other side-chains in the protein structure. We therefore need a procedure under which all the involved side-chains along the backbone of the protein can relax in a direction towards a new equilibrium position.

According to the mechanism just described, the side-chain of a given level is correlated with the side-chains of the preceding levels. A level is here defined by all the preceding side-chains, in a given succession of side-chains. If, however, the relaxation was just proceeding sequentially along the backbone, the underlying dynamics would not have any physical reality.

We shall instead consider a relaxation proceeding sequentially through the present succession of neighbors in the three-dimensional configuration space. (For a coordination number of two, the succession of neighbors is uniquely defined once a direction is fixed.) It is important that the relaxation of the side-chain motion is non-exponential in time, which observations [10] indicate, and which has been found to occur in systems with a hierarchically constrained dynamics [20]. The relaxation follows a power law, since the weights in each level are constant.

The best *sequence* of levels, i.e. the most reasonable choice of neighbors, is defined so as to make the total interaction energy of the entire protein minimal. This is achieved by spanning out a travelling salesman tour through points of attack in all the involved side-chains. The sequence of side-chains in the tour is selected by optimizing the interaction energy, which is a function of the euclidian distances between the side-chains. The procedure thus encompasses both interactions of short range as well as fewer interactions of a longer range.

3.2 The potential

Instead of optimizing the bare tour length, as in the ordinary TSP, we shall optimize a "length" which is a sum of distances each transformed by a function f. The function f describes the resulting potentials of the interaction

between two side-chains.

We are using an approximation to the following well-known potential which includes polar as well as van der Waals interactions:

$$f(d) = \frac{A}{d^{12}} - \frac{B}{d^6} + \frac{q_1 q_2}{dD}, \tag{3.1}$$

where D is the dielectric constant of the medium, A and B are known constants taken from reference [21] and [22], and q_1 and q_2 are charges.

The precise form of the set of functions used to mimic this potential is shown in figure 1, where the two curves correspond to attractive and repulsive interactions, depending on the hydrophobicity index and polarity of the side-chains. The position of the first cusp on the curves is determined by van der Waals radii around the points of attack of the side-chains, while the position of the second cusp is determined by a cutoff which ensures that distant repulsive side-chains are unlikely to be chosen as neighbors.

The best nearest neighbor topology of the protein structure can thus be found by solving an ordinary Travelling Salesman Problem in which the distances between cities are the values of the function f instead of the euclidian distances.

3.3 The procedure

In detail our procedure is as follows. Consider a given protein in the tertiary structure, where all the coordinates along the folded backbone are known from x-ray crystallography or NMR. The side-chains of all the amino acids along the backbone move according to their mutual interaction, and under constraints from the bonding geometry. We now lay out cones along the backbone, one for each side-chain, placed with their tops at the carbon-β atoms and with axes along the carbon-α—carbon-β direction, as shown in figure 2. These local cone "state spaces" are discretized and finite, and they emerge when the dihedral angles of the side-chains are varied. In the case of alanine, for example, the state space degenerates to a single point.

At the start, a random "travelling tour" between randomly chosen positions on the cones is generated. Then the sequential relaxation for this permutation of the cones takes place, where the new side-chain configuration on one cone is determined by the relaxation condition that the interaction energy, or (transformed) distance to the side-chain of the preceding cone, is minimized. When the sequential relaxation is completed a total tour length L can be calculated, and it represents the energy of the system.

The random starting tour represents the present set of "active" nearest-neighbor interactions, most of which surely are totally unphysical, and thus, by the potential in figure 1, has a rather high energy. By introducing a 2-bond move into this tour, another one is produced, and the same relaxation procedure is carried out sequentially, resulting in a new total energy of the system. The new tour and the resulting positions on the cones are kept as the new configuration, or rejected, according to the Boltzmann factor in

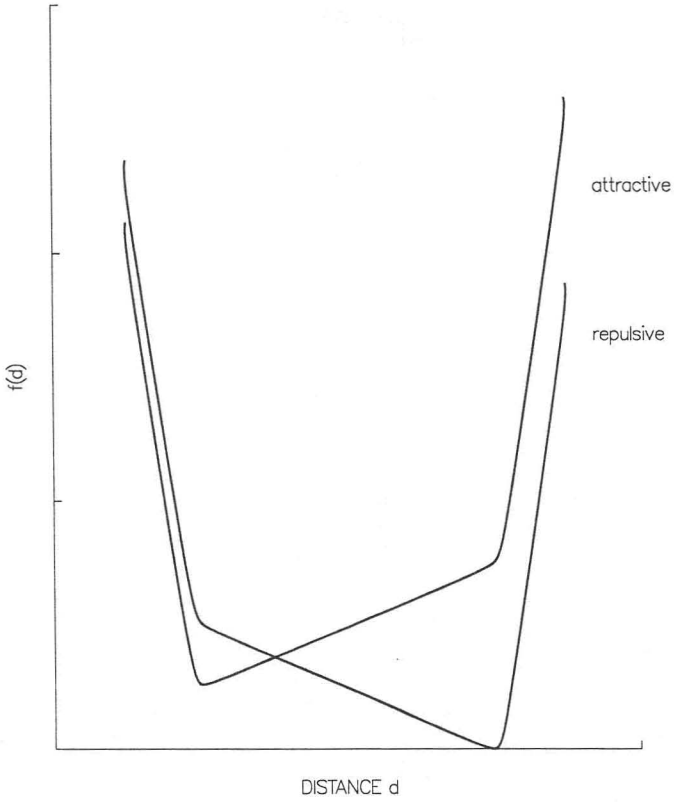

Figure 1: The new distances (or energies) as a function of the euclidian distance between the side-chains. The two curves correspond to attractive and repulsive interactions respectively. By assigning large values to both very short and very long distances, neighbors of this kind become highly improbable.

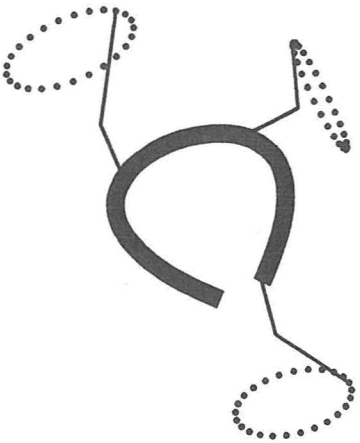

Figure 2: The degrees of freedom for three side-chains drawn as grids on cones.

the annealing scheme. We gradually cool down the protein system, accepting fewer and fewer changes in the tour while optimizing the length of the "round trip" between the side-chains of the protein. Meanwhile the protein is relaxing into a conformational state or substate.

The resulting substate will in general first of all depend on the location of the cones, i.e. the initial backbone structure. These substates are local or global 2-opt minima of the tour length (2.1), and they are in the next section compared with the native structure obtained by x-ray data. The difference between the native structure and the generated substates reflects the diversity of the nearest neighbor interactions activated by the protein.

It is important to notice that our procedure of finding local minima or conformational substates only depends on a few free parameters such as initial temperature and cooling velocity. The intrinsic parameters needed to fix the cones, such as cone size, polarity and angular constraints, are all data that are easy to obtain and in fact data that, except for the backbone geometry, follow from the sequence of amino acids.

3.4 Functional complexity

The set of all conformational substates possesed by a given protein determines the *functional complexity* of the structure. This set is, by our travelling salesman approach, described by a set of tours, in which each member represents a discrete characterisation of the interactions responsible for a particular conformational substate. In the appendix, it is shown how one

can easily define a metric on the set of all possible tours, and hence measure the distance between the interactions employed by different conformational substates. Armed with the inter-state distances it becomes possible to reveal the organization in the full set of active interactions, possibly that of a hierarchy [9,12,14]. A tree-like organization can be constructed by a classification process [23]. The classification process lumps the conformational substates, in groups within groups, according to their similarities. A set with an organization neither completely ordered nor completely disordered will correspond to a structure with a high diversity in the internal interactions, and hence a high complexity, while a low complexity will result from a set with random or self-similar organization. Our approach thus makes it possible to compare protein structures with respect to functional complexity.

4. Numerical results from computer simulations

We shall in this section illustrate our approach to protein conformations by two examples: (1) Avian Pancreatic Polypeptide, APP, a medium-size protein of 36 amino acids, and (2) Leucin Enkephalin, a very small polypeptide of only 5 amino acids.

We choose these two simple examples of polypeptides for this study because their side-chain motions are easy to follow in detail, and are well documented in the literature [24–30].

4.1 Avian Pancreatic Polypeptide (APP)

The first example, APP, is interesting from a modelling point of view, due to its compact globular structure comprising an alpha-helix (residues 14–32) and a hydrophobic core, and yet being quite simple compared to ordinary globular proteins. Its crystalline structure is well known from x-ray diffraction [24], and is shown in figure 3a,b.

We simulate the APP by fixing the center of our global coordinates at the carbon-α atoms of each amino acid on the backbone. We allow the side-chains to move on 'grid' points on cones as described above, representing possible configurations of the side-chains of each amino acid. Optimal or near-optimal positions on the cones are to be determined in the simulation. Due to the frustrated situation with both positive and negative interactions many different substate configurations must be expected to have low cost in interaction energy.

The optimization procedure starts with the side-chains in random positions on the cones and a random choice for the nearest neighbor topology. The corresponding tour is crossing in and out through the entire molecule, as shown in figure 4a. When the annealing proceeds the tour becomes stable at a certain temperature and all the side-chains have relaxed accordingly. The result of the optimization is shown in figure 4b, and represents a set of interactions with a much lower energy. The rate at which the temperature was lowered is logarithmic [18].

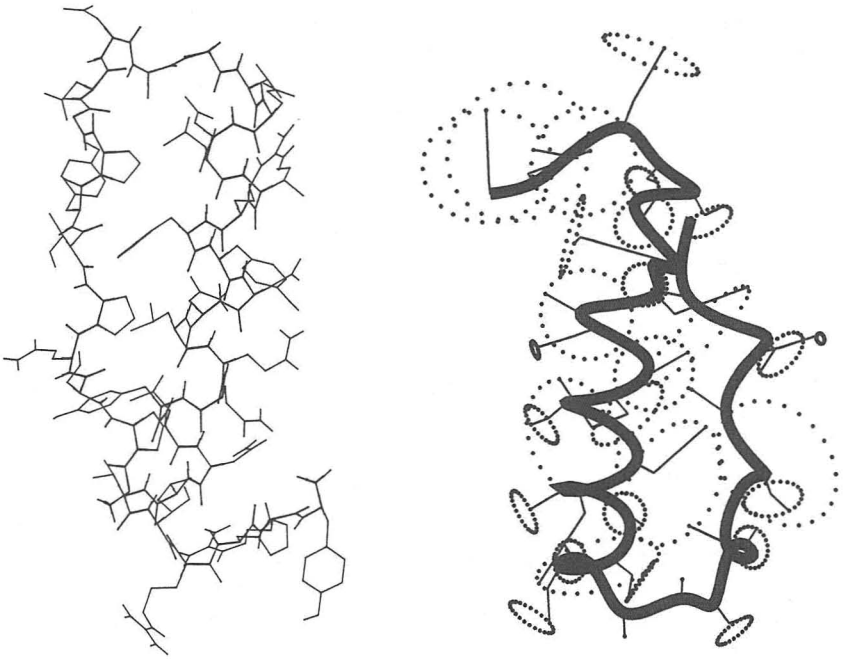

Figure 3: (a) The x-ray structure of Avian Pancreatic Polypeptide (APP) a medium-size protein of 36 amino acids. (b) APP with the cone state-spaces for each side-chain.

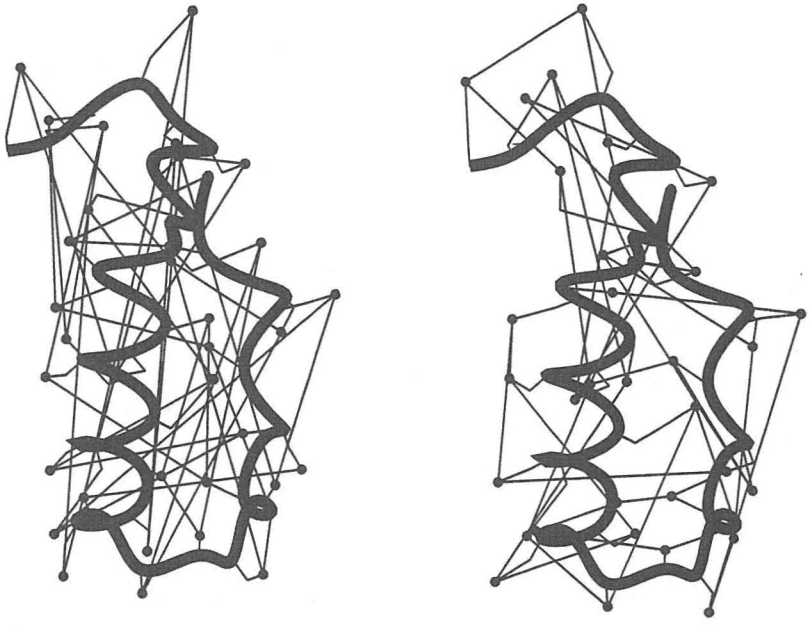

Figure 4: (a) A random starting configuration for Avian Pancreatic Polypeptide (APP). Most of the interactions are clearly unphysical as nearest neighbor interactions. $L_{\text{start}} = 7320.1$. (b) A self-consistent set of nearest neighbors, $L_{\text{end}} = 2264.3$. The L's are given in transformed distances representing interaction energy.

Our results can be compared with known conformational changes [25,26]. The fairly fixed three-dimensional structure due to a stable hydrophobic core is presumably essential for a correct binding to the hormone receptor. Especially the direction of Tyr 27 and Phe 20 bending away from the alpha-helix towards the proline chain is thought to be crucial for the stability of the functional state of APP. Asp 22 and Glu 25 should point in the opposite direction away from the proline chain, and thus together constitute hydrophobic and hydrophilic sides of the alpha-helix. Our data in the minimized configuration of figure 4b describe that picture.

In Table 1 is shown RMSD, standard-deviations and χ-square for the sample of minimized conformations in relation to x-ray data. We found RMSD values of the order of 1.6–2.7 Å. Within an average level of standard-deviations of the order of 2.1 Å there is statistical justification, for infering that the data is consistent with the x-ray structure. The full sample being the result of one hundred runs is shown in figure 5.

One could question the importance of the comparison between the sample of substates and x-ray data from the crystalline structure of the protein. This structure is presumably an artifact of the conditions for crystal growth. However, in default of better data for the biologically active protein, e.g. NMR data for the protein in solution, it is the best we can do at the moment. It is important to note that our method, in principle, makes it possible to take into account the effect of the surrounding media by simply including solvent molecules as cities in the TSP approach.

4.2 Leu_5-enkephalin

The second example of (Leu_5)-enkephalin (Tyr-Gly-Gly-Phe-Leu) has only five side-chains. The three large and important side-chains, Tyr, Phe and Leu are represented by big cones sticking out from the corresponding carbon-β atoms, and their motions are essential to the various conformations. The unfolded form has been observed to participate in beta-sheet polymers, reference [27], but we shall study its folded structure, see figure 6a,b and reference [28], stabilized by two hydrogen bonds between N1, O4 and O1, N4. The two glycine amino acids are kept fixed during the simulation, but they are still interacting with other side-chains, especially due to their repulsive cores.

Typical results of a computer simulation are shown in figure 7a,b, start and end configurations respectively. The end results show a fairly fixed orientation of the two important side-chains, Phe and Leu, while the Tyr side-chain is free to move almost perpendicular to the ring plane. This is what has been observed in x-ray diffraction data [28–30].

Smaller peptides are known to be less constrained than larger ones. Compared to their size, they are employing a more diverse set of interactions than the densely packed globular structures. The mean-deviation here is higher than in the case of APP, and is within 3.4 Å of RMSD. The standard-deviations are listed in Table 2, and they lie in the interval of 1.3 and 2.8 Å.

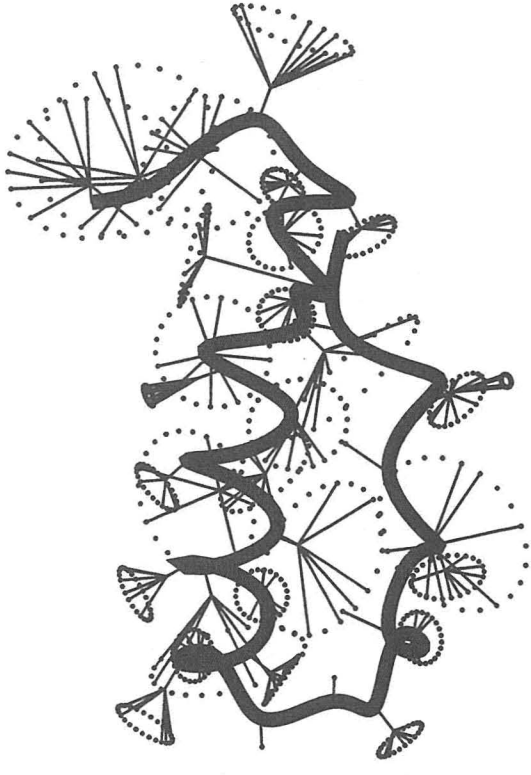

Figure 5: The sample of one hundred runs on APP is shown with their distribution on sites in the cone state spaces. The weight of each site is not shown.

i	SD	MD	$MAXD$
1	0.0	0.0	0.0
2	0.0	0.0	0.0
3	1.3	1.1	2.5
4	0.3	0.2	0.6
5	0.0	0.0	0.0
6	1.2	0.6	3.0
7	4.0	3.7	7.9
8	0.0	0.0	0.0
9	0.0	0.0	0.0
10	1.4	1.5	2.7
11	1.4	1.4	2.8
12	0.0	0.0	0.0
13	0.0	0.0	0.0
14	1.4	1.2	2.7
15	1.4	1.2	3.0
16	1.4	1.7	2.8
17	1.1	0.5	2.8
18	1.3	2.2	3.0
19	2.9	2.1	6.8
20	3.8	4.2	7.9
21	3.9	3.8	7.7
22	1.3	1.5	2.7
23	2.2	1.6	4.8
24	2.2	3.5	4.9
25	0.6	0.6	1.1
26	2.5	2.2	5.0
27	3.5	5.9	7.8
28	1.4	1.3	2.9
29	2.3	2.1	4.7
30	1.4	1.0	3.1
31	1.3	0.9	2.9
32	1.1	1.4	2.3
33	2.7	1.9	5.6
34	3.0	2.1	6.7
35	3.7	4.0	7.5
36	3.9	3.8	8.2

Table 1: APP. i: Amino acid number, SD: Standard deviation, MD: Mean deviation from x-ray structure, $MAXD$: Maximum deviation from x-ray structure in the cones. $\chi^2 = 0.77$.

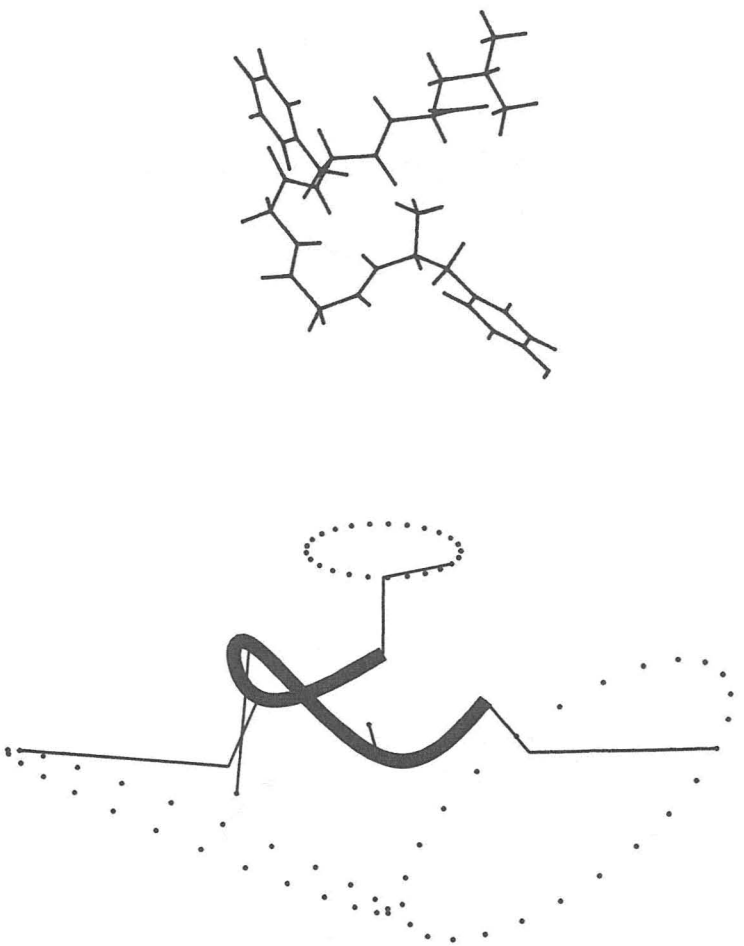

Figure 6: (a) The x-ray structure of Leu_5-enkephalin, a small peptide of 5 amino acids. (b) Leu_5-enkephalin with the cone state-spaces for each side-chain.

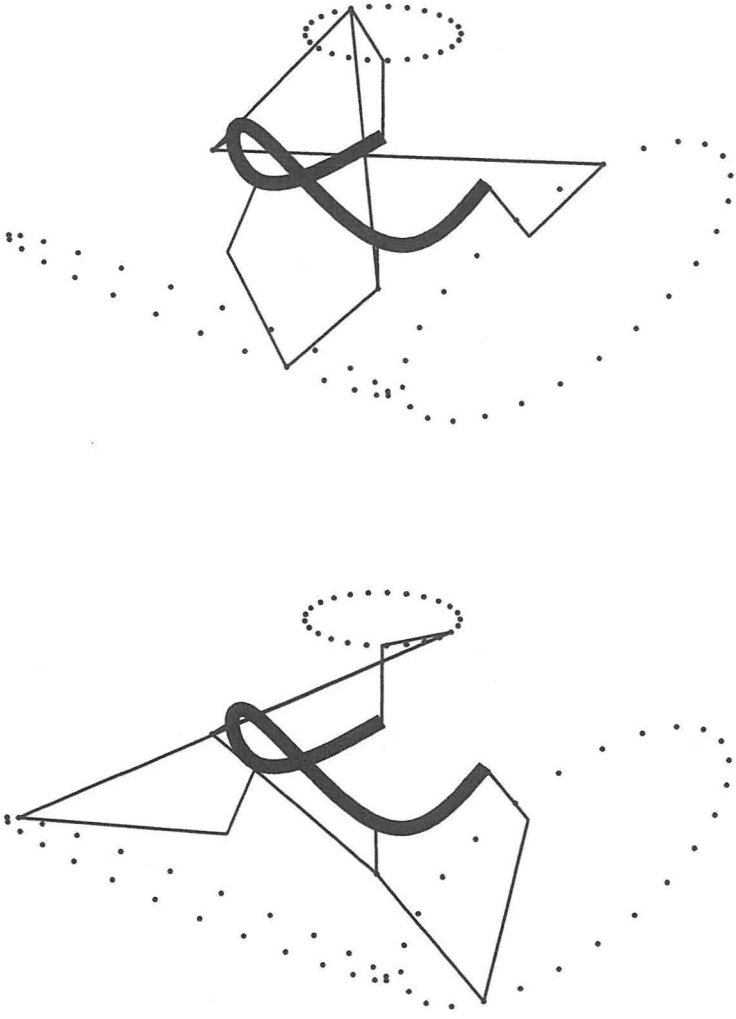

Figure 7: (a) The random starting configuration for Leu_5-enkephalin, $L_{\text{start}} = 372.2$. (b) Typical end configuration, $L_{\text{end}} = 131.9$.

i	SD	MD	$MAXD$
1	2.8	6.8	7.9
2	0.0	0.0	0.0
3	0.0	0.0	0.0
4	1.3	0.4	7.9
5	1.4	1.1	2.8

Table 2: Leu_5-enkephalin. i: Amino acid number, SD: Standard deviation, MD: Mean deviation from x-ray structure, $MAXD$: Maximum deviation from x-ray structure in the cones. $\chi^2 = 3.8$.

The full sample is shown in figure 8.

5. Conclusion

We have presented a method for finding conformational substates of proteins which relies on an analogy to a hard well-known problem from the field of optimization, the travelling salesman problem. Efficient techniques for obtaining near-optimal solutions for this problem have provided us with a tool for finding conformational substates, a tool which drastically decreases the computational complexity that usually haunts protein engineers. The model gives only crude, but important, features of the substate ensembles, and it contains only a few parameters to be adjusted in the search of local minima or conformational substates of the protein structure. The necessary parameters are easily obtained from x-ray data of the backbone coordinates, and from the amino acid sequence, such as hydrophobicity, polarity and size of individual amino acids.

A quantitative complexity measure characterizing the functional complexity of a protein structure has furthermore been proposed. It is our intention to apply the complexity measure to available protein data, and to extend the model to optimize not only local side-chain coordinates but also global backbone coordinates, i.e. to treat the substate conformations as well as the folding of proteins. This extension is natural, since the folding can be considered as being a relaxation process, which dynamically minimizes the backbone interaction energy, according to a changing neighbor topology, exactly as in the TSP approach.

Acknowledgments

We wish to thank J. Bohr, R.M.J. Cotterill, C.S. Jacobsen, J. Hertz, M. Kielland-Brandt, E. Mosekilde, O.G. Mouritsen, L. Nørskov, O.H. Olsen, S. Petersen, T. Schwartz, T. Særmark, for valuable discussions.

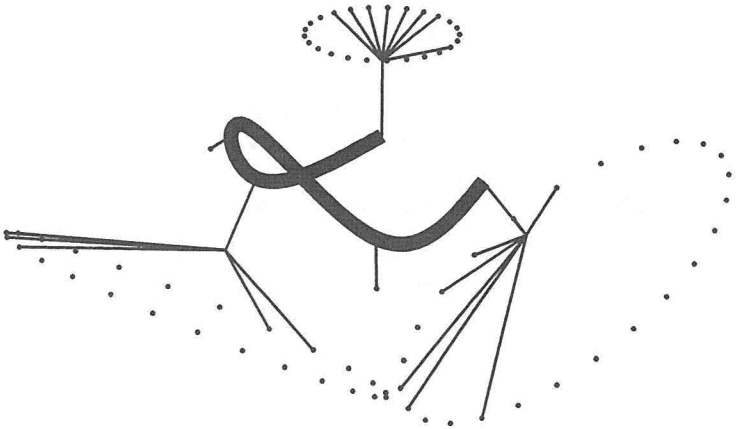

Figure 8: The sample of one hundred runs on Leu_5-enkephalin is shown with their distribution on sites in the cone state spaces. The weight of each site is not shown.

Appendix A. Appendix

The interactions participating in a conformational substate are, by our method, described by a tour, which is a unique selection of N links from the full set of $N(N-1)/2$ possible TSP links. Such a selection will be a point in the space of allowed tours. Distances between points in this space can be defined in various ways, either as the number of 2-bond moves which turn one tour into another [31], or simply as the fraction of links that differ in two tours

$$d_{12} = 1 - \frac{1}{N} \sum_i l_i^1 l_i^2. \tag{A.1}$$

The overall sum is $N(N-1)/2$ links, and $l_i^m = 1$ if the link is participating in tour m, and 0 if it is not. Hence $0 \leq d_{12} \leq 1$. Correspondingly, an overlap $q_{12} = 1 - d_{12}$, between two conformational substates can be defined as the fraction of links they have in common, $0 \leq q_{12} \leq 1$.

In tour-space the solutions will not be randomly distributed, but will be lumped in some way. When the tours are subjected to a classification process [23], the lumps can be characterized by a tree, showing the distances between clusters and the actual number of them. (The classification process redefines distances between tours, but makes, according to some criteria, as few modifications as possible.)

A measure of complexity [12] can be assigned to this tree (hierarchy) of functional states by counting the number of distinct subtrees the tree possesses, which again characterizes the diversity of the interactions present in

the protein structure. A highly uniform tree will correspond to few distinct interactions, as will a random one. In contrast a non-uniform but non-random tree will correspond to interactions which are highly heterogeneous. The complexity assigned to a protein structure will thus be related to the functional behavior of the protein, and not to a description of its primary, secondary or tertiary structure. It would be very interesting, however, to compare this complexity measure with the measure of the information-content obtained from a sequence analysis.

The complexity measure can be normalized so that different proteins can be compared with respect to functional complexity.

Without a discrete characterization of the conformational substates, as our tours are, it is hard to see how a metric can be defined. And without a metric, no classification, and hence no tree, can represent the set of substates for a given protein in an absolute way.

References

[1] S. Wolfram, "Approaches to Complexity Engineering," *Physica* **22D** (1986) 385.

[2] R.H. Austin, K.W. Beeson, L. Eisenstein, H. Frauenfelder, and I.C. Gunsalus, *Biochemistry* **14** (1975) 5355.

[3] H. Frauenfelder in *Structure and Dynamics: Nucleic Acids and Proteins*, eds. E. Clementi and R.H. Sarma, (Adenine, Guilderland, NY, 1983) 369.

[4] A.T. Brünger, G.M. Clore, A.M. Gronenborn, M. Karplus, *Proc. Natl. Acad. Sci.* **83** (1986) 3801.

[5] J.A. McCammon and M. Karplus, *Ann. Rev. Phys. Chem.* **31** (1980) 29.

[6] J. Åqvist, W.F. van Gunsteren, M. Leijonmarck and O. Tapia, *J. Mol. Biol.* **183** (1985) 461.

[7] M. Levitt, C. Sander, P.S. Stern, *J. Mol. Biol.* **181** (1985) 423.

[8] G. Neámethy and H.A. Sheraga, *Q. Rev. Biofys.* **10** (1977) 239.

[9] A. Ansani, J. Berendzen, S.F. Bowne, H. Frauenfelder, I.E.T. Iben, T.B. Sauke, E. Shyamsunder and R.D. Young, *Proc. Natl. Sci. USA* **82** (1985) 5000 (and references therein).

[10] D.L. Stein, *Proc. Natl. Acad. Sci. USA* **82** (1985) 3670.

[11] D. Sherrington, "Disorder, Frustration and Metastability: The development of a new Era", in *Proceedings of 1986 Heidelberg Colloquium on Glassy Dynamics and Optimization.*

[12] T. Hogg and B.A. Huberman, "Complexity and Adaptation," *Physica* **22D** (1986) 376–384.

[13] S. Kirkpatrick and G. Toulouse, *J. Physique* **46** (1985) 1277.

[14] R. Rammal, G. Toulouse and M.A. Virasoro, *Rev. Mod. Phys.* **58:3** (1986) 765 (and references therein).

[15] C.H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization*, (Prentice-Hall, 1982).

[16] M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, (W.H. Freeman and Co., San Francisco, 1979).

[17] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, *Science* **220:N4598** (1983) 671.

[18] E.H.L. Aarts and P.J.M. van Laarhoven, *Simulated Annealing: A Review of the Theory and Applications*, (Kluwer Academic Publishers, 1987).

[19] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller, *J. Chem. Phys.* **21** (1953) 1087.

[20] R.G. Palmer, D.L. Stein, E. Abrahams and P.W. Anderson, *Phys. Rev. Lett.* **53:10** (1984) 958.

[21] B.R. Gelin and N. Karplus, *Biochemistry* **18** (1979) 1256.

[22] P.C. Gehlen, J.R. Beeler and R.I. Jaffee, Jr., *Interatomic Potentials and Simulation of Lattice Defects*, (Plenum Press, NY, 1972).

[23] A.D. Gordon, *Classification*, (Chapman and Hall, London, 1981).

[24] T.L. Blundell, J.E. Pitts, I.J. Tickle, S.P. Wood, and C.W. Wu, *Proc. Natl. Acad. Sci. USA* **78** (1981) 4175.

[25] T.W. Schwarts, *Pancreatic Polypeptide, a hormone under Vagal Control* (1983), published by the American Gastroenterological Association.

[26] I. Glover, I. Haneef, J. Pitts, S. Wood, D. Moss, I. Tickle, T. Blundell, *Biopolymers* **22** (1983) 293.

[27] I.L. Karle, J. Karle, D. Mastropaolo, A. Camerman, and N. Camerman, *Acta Cryst.* **B39** (1983) 625.

[28] G. David Smith and J.F. Griffin, *Science* **199** (1978) 1214.

[29] M.A. Khatet, M.M. Long, W.D. Thompson, R.J. Bradley, G.B. Brown, and D.W. Urry, *Bioch. & Biophys. Res. Com.* **76** (1977) 224.

[30] V. Isogai, G. Ńeámethy, H.A. Scheraga, *Proc. Natl. Acad. Sci. USA* **74** (1977) 414.

[31] S.A. Solla, G.B. Sorkin, S.R. White, in *Disordered Systems and Biological Organisation*, ed. E. Bienenstock, F. Fogelman Soulie, and G. Weisbuch, Nato ASI Series, F, **20** (1986) 283.